Journal Club

Editor's Note: These short, critical reviews of recent papers in the *Journal*, written exclusively by graduate students or postdoctoral fellows, are intended to summarize the important findings of the paper and provide additional insight and commentary. For more information on the format and purpose of the Journal Club, please see http://www.jneurosci.org/misc/ifa_features.shtml.

The Neural Mechanisms of Bayesian Belief Updating

Daniel Bennett
School of Psychological Sciences, The University of Melbourne, Parkville, Victoria 3010, Australia

Review of Vossel et al.

A central function of the nervous system is to use sensory information to infer the causal structure of the external world. According to Bayes' rule, the optimal way of using this information is to calculate the information's likelihood under various models of the environment, and to weight this likelihood by the strength of prior belief in each model to derive posterior beliefs. In recent years, the influential hypothesis has been advanced that Bayesian inference represents a unifying principle of neural computation (the Bayesian brain hypothesis; Knill and Pouget, 2004). This framework has been applied to many topics, including vision, motor planning, and behavioral conditioning (Courville et al., 2006; Körding and Wolpert, 2006; Yuille and Kersten, 2006), with the overarching goal of identifying how neural computations implement optimal Bayesian statistical principles.

One particular focus of this research has been Bayesian belief updating: the transformation of prior beliefs into posterior beliefs when new information is observed. Although functional magnetic resonance imaging (fMRI) and electroencephalography have been used to identify some of the regions and processes involved in belief updating (O'Reilly et al., 2013; Bennett et al., 2015; Kolossa et al., 2015), an important open question is how these regions interact to update beliefs. Addressing this question would represent a step beyond localization toward a process-based account of the neural mechanisms of Bayesian belief updating.

In a recent paper published in *The Journal of Neuroscience*, Vossel et al. (2015) investigated this question using fMRI data acquired from 18 healthy participants completing a Posner cueing task (Posner, 1980). In this task, participants made visual saccades toward a target—a suprathreshold Gabor patch—which appeared either left or right of a central fixation point. Before the target was displayed, the presentation of a left- or right-pointing cue predicted the target location with varying cue validity (CV), defined as the proportion of cues that correctly predicted the target. Every 32–36 trials throughout the experiment, CV was changed pseudorandomly to one of three levels (88%, 69%, 50%). Crucially, participants were not aware in advance of when or how CV would change, and so had to learn these contingencies on the basis of experience: a Bayesian inference problem.

To characterize trial-by-trial belief updating, Vossel et al. (2015) analyzed behavioral data using a Bayesian computational model, the hierarchical Gaussian filter (HGF; Mathys et al., 2011), which has previously been successfully applied to Posner cueing tasks (Vossel et al., 2014). The HGF used response speed (RS) as a dependent measure, and assumed that RSs were generated by a bipartite system consisting of a perceptual model, which tracked beliefs about changing CVs, and a response model, which mapped belief strength onto RSs. The HGF's perceptual model took the form of a three-level Gaussian hierarchy. At the lowest level of this hierarchy were participants' observations of validly and invalidly cued targets, assumed to be generated by a latent trial-specific CV. At the second level of the HGF hierarchy, CV evolved across trials as a Gaussian random walk to capture the fact that CVs were nonstationary. At the third level of the HGF, the volatility of the second-level random walk itself changed over time as a Gaussian random walk, capturing the fact that participants experienced periods of both stability and volatility during the task. Second- and third-level random walks were respectively parameterized by participant-specific precision parameters ω and ϑ, which were assumed to drive individual differences in behavior. For subsequent fMRI analyses, a key feature of the HGF was that on each trial, it estimated the attention-weighted precision of beliefs, denoted $\alpha(\hat{\pi}^{(t)}{}_{1})$, which can be interpreted as the proportion of attentional resources allocated to the cued location on a given trial.

Vossel et al. (2015) used these participant- and trial-specific values of $\alpha(\hat{\pi}^{(t)}{}_{1})$ to identify brain regions associated with Bayesian belief updating. fMRI data were first analyzed using a general linear model (GLM) with four first-level regressors of interest: valid, invalid, left-

Received Oct. 13, 2015; revised Nov. 12, 2015; accepted Nov. 17, 2015.

I thank Stefan Bode, Bowen Fung, Carsten Murawski, Daniel Rosenblatt, Christina van Heer, Katharina Voigt, and Hayley Warren for helpful comments and feedback on the draft manuscript.

The author declares no competing financial interests.

Correspondence should be addressed to Daniel Bennett, School of Psychological Sciences, The University of Melbourne, Parkville, VIC 3010, Australia. E-mail: daniel.bennett@unimelb.edu.au.

DOI:10.1523/JNEUROSCI.3742-15.2015

0270-6474/15/3516300-03$15.00/0

ward, and rightward cues. Trial-specific values of $\alpha(\hat{\pi}^{(t)}{}_1)$ were then extracted from the HGF and used to parametrically modulate each of these four regressors. Next, a second-level analysis located brain regions where cue validity interacted with the parametric effect of $\alpha(\hat{\pi}^{(t)}{}_1)$ [indicating different $\alpha(\hat{\pi}^{(t)}{}_1)$ regression slopes for valid vs invalid cues]. This pattern was taken to be indicative of a region's involvement in Bayesian belief updating, since invalid cues violate expectancies and signal a potential change in CV, and should therefore trigger larger belief updates than valid cues. Moreover, this relationship should be modulated by attention-weighted precision: when $\alpha(\hat{\pi}^{(t)}{}_1)$ is high, as in a block with high CV, an invalid cue is a stronger belief-updating trigger than when $\alpha(\hat{\pi}^{(t)}{}_1)$ is lower. Using these criteria, the second-level GLM analysis identified three regions involved in Bayesian belief updating: right anterior putamen, right frontal eye fields (FEF), and right temporo-parietal junction (TPJ).

Next, to investigate how Bayesian belief updating was implemented, these three areas were designated as regions of interest for dynamic causal modeling (DCM) analysis. It was found that the best-fitting DCM model was one in which stronger beliefs [higher values of $\alpha(\hat{\pi}^{(t)}{}_1)$] were associated with decreased correlation between TPJ and FEF activity for validly cued trials, and with increased correlation between TPJ and FEF activity on invalidly cued trials (Vossel et al., 2015, their Fig. 6). This was taken to indicate a coupling between ventral (TPJ) and dorsal (FEF) streams of visual processing, suggesting that coordinated activity between TPJ and FEF may reflect transmission of a belief-updating signal. Such a signal could upregulate attention-related dorsal stream activity when expectancies were violated following invalid cues, and downregulate activity following valid cues. This conclusion is consistent with, and provides a mechanistic explanation for, previous work implicating TPJ in Bayesian updating of internal models of the environment (Geng and Vossel, 2013).

Although anterior cingulate cortex (ACC) has been implicated in Bayesian belief updating in previous research (O'Reilly et al., 2013), ACC was not one of the belief-updating regions identified by Vossel et al. (2015). This inconsistency may be driven by use of different belief-updating metrics in different studies: rather than define belief updates by an interaction between CV and $\alpha(\hat{\pi}^{(t)}{}_1)$, O'Reilly et al. (2013) searched for regions encoding belief-update magnitude. These different metrics might therefore have identified distinct components of a broader-scale belief-updating network. Similarly, although ACC is thought to encode environmental volatility in learning under uncertainty (Behrens et al., 2007), Vossel et al. (2015) found no significant effect of volatility in ACC or any other brain region. One possible explanation for this discrepancy is the Posner cueing task's very short cue–target interval (600 ms). Because of the poor temporal resolution of fMRI, this would have meant that the GLM analysis was unable to disentangle cue processing from response execution. As a result, the fMRI results of Vossel et al. (2015) may not be directly comparable to past research by Behrens et al. (2007). Future studies could disentangle these discrete task stages using a longer cue–target interval or by using a technique with better temporal resolution, such as electroencephalography.

A particular strength of the study by Vossel et al. (2015) was the manner in which it combined Bayesian computational modeling with fMRI data analysis. For both behavioral data and GLM analysis, the authors were able to show that the HGF fit data better than two non-Bayesian competitor models: a Rescorla-Wagner learning rule and a model assuming participants knew the true CV in each block. This supports the conclusion that participants were behaving Bayes-optimally. However, an important caveat here is that the flexibility afforded to some Bayesian observer models by their parameterization and choice of prior might mean that it is not possible to empirically falsify the hypothesis that the brain acts as a Bayesian observer (Daunizeau et al., 2010; Bowers and Davis, 2012). Although it is strongly suggestive that a Bayesian model explained both behavioral and neural data better than non-Bayesian competitors, this is not logically sufficient to show that participants necessarily acted as Bayesian observers. Indeed, Vossel et al. (2015) do not make this claim.

Moreover, it remains unclear whether the assumption of Bayes-optimality in the HGF is viable in more complex environments than a Posner cueing task. Payzan-LeNestour and Bossaerts (2011) demonstrated that in complex environments, Bayesian models became increasingly computationally intractable, and no longer fit behavioral data better than non-Bayesian competitors. Furthermore, even in environments suitable for Bayesian inference, simple heuristics can provide a better account of behavior for considerable subsets of participants (Steyvers et al., 2009; Bennett et al., 2015). A potential solution to this problem is given by a recent study showing that constraints based in principles of efficient sensory coding enabled a Bayesian model to explain seemingly anti-Bayesian percepts (Wei and Stocker, 2015). Similarly, constraining Bayesian observer models by neurophysiological principles such as capacity limits on processing may allow these models to be successfully applied to more complex environments.

In summary, the work of Vossel et al. (2015) provides a compelling synthesis of behavioral modeling and neuroimaging. By combining a sophisticated behavioral model with DCM analysis of neural data, the authors identified potential neural mechanisms of Bayesian belief updating in deployment of spatial attention. These findings represent a valuable step toward a process-based account of belief updating in the Bayesian brain.

References

Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. Nat Neurosci 10:1214–1221. CrossRef Medline

Bennett D, Murawski C, Bode S (2015) Single trial event-related potential correlates of belief updating. eNeuro 2:e0076-15.2015 1–14. CrossRef Medline

Bowers JS, Davis CJ (2012) Bayesian just-so stories in psychology and neuroscience. Psychol Bull 138:389–414. CrossRef Medline

Courville AC, Daw ND, Touretzky DS (2006) Bayesian theories of conditioning in a changing world. Trends Cogn Sci 10:294–300. CrossRef Medline

Daunizeau J, den Ouden HE, Pessiglione M, Kiebel SJ, Stephan KE, Friston KJ (2010) Observing the observer (I): meta-Bayesian models of learning and decision-making. PLoS One 5:e15554. CrossRef Medline

Geng JJ, Vossel S (2013) Re-evaluating the role of TPJ in attentional control: contextual updating? Neurosci Biobehav Rev 37:2608–2620. CrossRef Medline

Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27:712–719. CrossRef Medline

Kolossa A, Kopp B, Fingscheidt T (2015) A computational analysis of the neural bases of Bayesian inference. Neuroimage 106:222–237. CrossRef Medline

Körding KP, Wolpert DM (2006) Bayesian decision theory in sensorimotor control. Trends Cogn Sci 10:319–326. CrossRef Medline

Mathys C, Daunizeau J, Friston KJ, Stephan KE (2011) A Bayesian foundation for individual learning under uncertainty. Front Hum Neurosci 5:39. CrossRef Medline

O'Reilly JX, Schüffelgen U, Cuell SF, Behrens TE, Mars RB, Rushworth MF (2013) Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. Proc Natl Acad Sci U S A 110:E3660 –E3669. CrossRef Medline

Payzan-LeNestour E, Bossaerts P (2011) Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. PLoS Comput Biol 7:e1001048. CrossRef Medline

Posner MI (1980) Orienting of attention. Q J Exp Psychol 32:3–25. CrossRef Medline

Steyvers M, Lee MD, Wagenmakers EJ (2009) A Bayesian analysis of human decision-making on bandit problems. J Math Psychol 53:168 –179. CrossRef

Vossel S, Mathys C, Daunizeau J, Bauer M, Driver J, Friston KJ, Stephan KE (2014) Spatial attention, precision, and Bayesian inference: a study of saccadic response speed. Cereb Cortex 24:1436 –1450. CrossRef Medline

Vossel S, Mathys C, Stephan KE, Friston KJ (2015) Cortical coupling reflects Bayesian belief updating in the deployment of spatial attention. J Neurosci 35:11532–11542. CrossRef Medline

Wei XX, Stocker AA (2015) A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. Nat Neurosci 18:1509 –1517. CrossRef Medline

Yuille A, Kersten D (2006) Vision as Bayesian inference: analysis by synthesis? Trends Cogn Sci 10:301–308. CrossRef Medline